NeuroImage 56 (2011) 140–148

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Technical Note

Extracting information from functional connectivity maps via function-on-scalar regression

Philip T. Reiss [a,b,*], Maarten Mennes [a], Eva Petkova [a,b], Lei Huang [a], Matthew J. Hoptman [b,c], Bharat B. Biswal [d,b], Stanley J. Colcombe [b], Xi-Nian Zuo [a,e], Michael P. Milham [a,b]

[a] *Department of Child and Adolescent Psychiatry, New York University School of Medicine, NY, USA*
[b] *Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY, USA*
[c] *Department of Psychiatry, New York University School of Medicine, NY, USA*
[d] *Department of Radiology, University of Medicine and Dentistry of New Jersey, Newark, USA*
[e] *Institute of Psychology, Chinese Academy of Science, Beijing, China*

ARTICLE INFO

Article history:
Received 23 November 2010
Revised 27 January 2011
Accepted 28 January 2011
Available online 4 February 2011

Keywords:
Functional connectivity
Functional data analysis
Model selection
Quantile regression
Resting state
Seed region

ABSTRACT

Functional connectivity of an individual human brain is often studied by acquiring a resting state functional magnetic resonance imaging scan, and mapping the correlation of each voxel's BOLD time series with that of a seed region. As large collections of such maps become available, including multisite data sets, there is an increasing need for ways to distill the information in these maps in a readily visualized form. Here we propose a two-step analytic strategy. First, we construct connectivity–distance profiles, which summarize the connectivity of each voxel in the brain as a function of distance from the seed, a functional relationship that has attracted much recent interest. Next, these profile functions are regressed on predictors of interest, whether categorical (e.g., acquisition site or diagnostic group) or continuous (e.g., age). This procedure can provide insight into the roles of multiple sources of variation, and detect large-scale patterns not easily available from conventional analyses. We illustrate the proposed methods with a resting state data set pooled across four imaging sites.

Introduction

The last several years have witnessed a surge of interest in studying brain connectivity by functional magnetic resonance imaging (fMRI) of the resting brain (e.g., Biswal et al., 1995; Damoiseaux et al., 2006; Margulies et al., 2007; Roy et al., 2009). Resting state functional connectivity (RSFC) is often represented by mapping the correlation of each voxel's BOLD time series with that of a pre-defined "seed" region of interest (ROI). With the growing availability of large collections of such maps, including multisite data sets (Biswal et al., 2010), there is an increasing need for ways to summarize such data and to visualize them at a glance—both for quality assurance and to understand different sources of variation.

Here we describe a novel technique to summarize three-dimensional seed-based RSFC maps into one-dimensional functions by collapsing them along a dimension representing physical *distance* from the seed, a quantity whose relationship with RSFC has been a topic of recent interest. For example, Kelly et al. (2009) chose five seed ROIs in the anterior cingulate cortex and examined how whole-brain connectivity with these seeds varied with age. Compared to children, adults were found to have more significantly correlated voxels at long distances from the seed ROI (see also Andrews-Hanna et al., 2007). In related work studying connectivity graphs among several dozen ROIs rather than connectivity of all voxels with a seed ROI, Fair et al. (2007, 2009) have proposed that the maturing brain progresses from "local" networks of short-range connections to a more "distributed" architecture incorporating longer-range connections. Thus, distance is a natural and informative dimension along which RSFC maps can be characterized.

Our proposed analytic strategy has two basic components. First, nonparametric quantile regression (Koenker, 2005) is used to derive subject-specific curves representing RSFC as a function of distance from the seed. We refer to these curves as "connectivity–distance profiles." Second, by viewing these profiles as "functional data objects" (Ramsay and Silverman, 2005), one can regress them on subject-specific predictors of interest using a fast new implementation of function-on-scalar regression (Reiss et al., 2010). Such regression analyses may reveal, for example, how short- or long-range connectivity develops with age, or how it differs between diagnostic groups.

We illustrate the utility of this strategy by analyzing a combined resting state data set acquired at four different labs, using six seed ROIs that have been well characterized in resting state fMRI studies

[*] Corresponding author at: Department of Child and Adolescent Psychiatry, New York University School of Medicine, 215 Lexington Ave., 16th floor, New York, NY 10016, USA. Fax: +1 212 263 2476.
E-mail address: phil.reiss@nyumc.org (P.T. Reiss).

1053-8119/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2011.01.071

(Fox et al., 2005). Connectivity–distance profiles with respect to the six seed ROIs provide strong visual evidence of systematic between-site differences, in line with the findings of Biswal et al. (2010). On the other hand, the age distributions for the four samples differ markedly, so that, in view of the above-cited evidence of age effects, the between-site differences could perhaps be due to age disparities. In other words, age and site may be acting as mutual confounders. Our proposed regression methodology offers a means to sort out the effects of age and site on RSFC.

Statistical methodology

Individual connectivity–distance profiles

We begin by explaining how to define, for an individual participant, a profile that captures the relationship between functional connectivity of a seed ROI and distance from that seed. Assume that there are V voxels in the brain and that, for $j = 1,\ldots,V$, the jth voxel lies at distance s_j from the center of the seed. (In what follows, we use simple Euclidean distance, although other distance measures such as geodesic distance can be substituted.) For each j, the RSFC c_j with the seed can be defined as the Pearson correlation of voxel j's BOLD signal during the scan with the BOLD signal for the seed region, following suitable preprocessing steps. The simplest reasonable definition of the connectivity–distance profile would be the estimated average connectivity as a function of distance from the seed. The estimate could be obtained by applying nonparametric regression (smoothing) methodology (e.g., Simonoff, 1996) to the scatterplot of connectivity vs. distance (as in the red curve in Fig. 1). This approach seeks to estimate a smooth function $y(s)$ such that

$$c_j = y\left(s_j\right) + \varepsilon_j \tag{1}$$

where ε_j is a random error term with expectation zero. The implicit simplifying assumption is that the expected value of c_j depends only on the distance of voxel j from the seed. Of course, in reality, two voxels at the same distance from the seed but in different directions may tend to differ systematically in RSFC due to various factors (e.g., they may belong to completely distinct brain structures or functional networks). However, for purposes of a straightforward exploratory analysis, we allow the distribution of ε_j to incorporate any such systematic effects, as well as truly random error.

The smooth profile function $y(s)$ given by Eq. (1) is the conditional expectation (mean) of c given distance s,

$$y(s) = E(c|s), \tag{2}$$

i.e., the average connectivity value on a sphere of radius s centered at the seed. However, scientific interest focuses less on average connectivity values than on finding regions whose connectivity with the seed is particularly strong (e.g., Kelly et al., 2009). This suggests that, instead of the conditional expectation [Eq. (2)], it is more useful to define the connectivity–distance profile by the conditional qth quantile $y_{(q)}(s) = F^{-1}(q|s)$, for a value q close to 1 (e.g., the upper green curve in Fig. 1). Here $F(\cdot|s)$ denotes the cumulative distribution function of the RSFC values conditional on distance s. The function $y_{(q)}$ can be estimated by nonparametric quantile regression methods (Koenker et al., 1994; Koenker, 2005). The chosen q should be high enough to capture extreme RSFC values, but not too high since the standard errors of the quantile estimates increase as q approaches 1. In what follows, we assume $q = 0.95$. Informal sensitivity analyses suggest that $q = 0.9$ or $q = 0.99$ results in very similar function estimates.

In addition to high-connectivity profiles based on the 95th percentile of RSFC as above, we can define low-connectivity profiles based on the 5th percentile (e.g., the lower green curve in Fig. 1). The latter might be useful when interest focuses on negative correlations between regions. In the analyses reported below, the low-connectivity profiles were typically negative except at short distances from the seed.

In summary, an individual's connectivity–distance profile captures the magnitude of RSFC with relatively strongly connected regions at any given distance from the seed. This representation of the relationship between distance and RSFC enables us to study between-group differences in short- and long-range connectivity, by means of the functional regression methodology described next.

Group analyses by function-on-scalar regression

The (high- or low-) connectivity profiles estimated for each of N participants, as described above, can be denoted by $\hat{y}_{(q)1}(s),\ldots,\hat{y}_{(q)N}(s)$. For simplicity, we shall treat them as known functions and suppress the dependence on q, resulting in the simpler notation $y_1(s),\ldots, y_N(s)$. Statisticians refer to data of this type as functional data, indicating that the observation for each participant consists of an entire function. Research on functional data analysis (Ramsay and Silverman, 2005) has produced functional-data analogs of many classical statistical

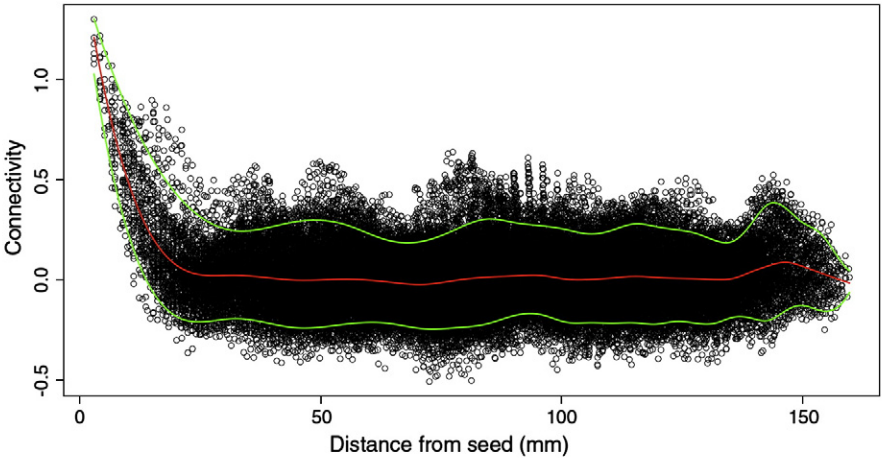

Fig. 1. Scatterplot of each voxel's connectivity with the lateral parietal seed of Fox et al. (2005) vs. distance from the seed, for a single participant. A nonparametric regression fit is shown in red, and nonparametric quantile regression fits (estimates of the 5th and 95th conditional percentiles) are shown in green.

techniques (see Viviani et al., 2005; Zhou et al., 2009, for previous applications to fMRI). In particular, multiple linear regression can be extended to the case of functional outcomes such as our connectivity–distance profiles (or other imaging-derived functions; see Zhu et al., 2010, for an application to diffusion tensor imaging tracts). The classical multiple regression model is given by

$$y_i = x_i^T \beta + \varepsilon_i, \tag{3}$$

where y_i is a scalar representing an outcome for the ith participant, x_i is a p-dimensional vector of predictor variables for that subject, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a vector of corresponding effects, and ε_i represents random error. A natural generalization of Eq. (3) to settings such as ours, in which each y_i is a function, is the functional linear model

$$y_i(s) = x_i^T \beta(s) + \varepsilon_i(s), \tag{4}$$

in which x_i is a vector of predictors as in Eq. (3), but the estimand $\beta = [\beta_1(\cdot), \ldots, \beta_p(\cdot)]^T$ is now a vector of functions (i.e., a vector-valued function) and the errors ε_i are functions. Two basic special cases of model (4) are of particular interest:

1. The one-way functional ANOVA model. Here the N participants are divided into G groups, such as different sites at which the resting state scans were acquired. If we denote by $g(i)$ the group to which participant i belongs, and let $x_i = (1, I_{g(i)=1}, \ldots, I_{g(i)=G})^T$ where $I_{g(i)=k}$ equals 1 if subject i belongs to group k and 0 otherwise, then, with a change of notation for the functional coefficient β, model (4) becomes $y_i = \mu(s) + \tau_{g(i)}(s) + \varepsilon_i(s)$. The functions $\tau_1, \ldots, \tau_G$ represent group effects which, for identifiability, are assumed to satisfy $\sum_{k=1}^{G} \tau_k(s) = 0$ for all s. As discussed below, this model can be used to test for group differences in the average profile.
2. Function-on-scalar regression with a single continuous predictor If $x_i = (1, x_i)^T$ for a continuous covariate x_i, say age, then model (4) can be written as

   $$y_i(s) = \beta_1(s) + x_i \beta_2(s) + \varepsilon_i(s). \tag{5}$$

   Then $\beta_2(\cdot)$ represents the effect of age, and for any x, the function $\hat{y}(\cdot) = \beta_1(\cdot) + x\beta_2(\cdot)$ represents the expected value of $y(\cdot)$ (i.e., the typical profile) conditional on age x. This model can be used to test for age effects.

As an aid to interpretation, it may be noted that if we took $\beta_2(s) = \beta_2 s$, the second term on the right side of model (5) would reduce to an ordinary interaction between x_i (age) and s (distance)—meaning that distance moderates the effect of age on connectivity, or alternatively that age moderates the distance-connectivity relationship. Model (5) goes beyond conventional moderation analysis in two ways: we allow the age effect to depend nonlinearly on s, and we extend the framework to modeling entire RSFC maps, by employing connectivity–distance profiles to capture the most salient information in these maps.

Estimating the functional linear model

To estimate the coefficient functions $\beta_1(\cdot), \ldots, \beta_p(\cdot)$, we posit that each lies in the span of a set of smooth basis functions $\theta_1(\cdot), \ldots, \theta_K(\cdot)$ defined on S, the range of distances of voxels from the seed. We choose a basis of cubic B-splines (see de Boor, 2001, for background). Since smoothness is controlled by the parameters $\lambda_1, \ldots, \lambda_p$ introduced below, the precise choice of the number of basis functions K is generally seen as much less critical, as long as it is large enough to capture the detail of the function(s) being estimated (Ruppert, 2002; Wood, 2006, p. 161); thus K is usually chosen informally, with values of 20–30 being typical. Restricting the coefficient functions in this way allows us to write $\beta(s) = B\theta(s)$ for some $p \times K$ matrix $B = (b_1 \ldots b_p)^T$, where $\theta(s) = [\theta_1(s), \ldots, \theta_K(s)]^T$. Fitting model (4) is thus reduced to estimating the pK elements of B. To do so, we minimize the penalized integrated sum of squared errors (SSE)

$$\sum_{i=1}^{N} \int_S \left[y_i(s) - x_i^T B\theta(s) \right]^2 ds + \sum_{m=1}^{p} \lambda_m b_m^T J b_m \tag{6}$$

where $\lambda_1, \ldots, \lambda_p$ are nonnegative constants, and J is the $K \times K$ matrix with (i, j) entry $\int_S \theta_i''(s)\theta_j''(s)ds$. The second term above equals $\sum_{m=1}^{p} \lambda_m \int_S \beta_m''(s)^2 ds$, and is known as a roughness penalty, since its effect is to discourage excessive roughness or wiggliness in the estimates of the functions β_m. The vector $\lambda = (\lambda_1, \ldots, \lambda_p)^T$ of smoothing parameters, which controls the extent to which roughness is penalized, can be chosen to minimize the "leave-one-function-out" cross-validation score (Rice and Silverman, 1991)

$$\sum_{i=1}^{N} \int_S \left[y_i(s) - \hat{y}_i^{(-i)}(s; \lambda) \right]^2 ds, \tag{7}$$

where $\hat{y}_i^{(-i)}(\cdot; \lambda)$ is the predicted value for the ith connectivity–distance profile, based on the model fitted to the other *N-1* profiles with smoothing parameters λ. In practice, it is often sufficient to use a common smoothing parameter $\lambda_1 = \ldots = \lambda_p = \lambda$ for all p coefficient functions. Traditionally, optimal smoothing by cross-validation has been computationally laborious for function-on-scalar regression (Ramsay et al., 2009, p. 154), but Reiss et al. (2010) show how the process can be made dramatically faster. Construction of confidence intervals for the coefficient functions is also described by Reiss et al. (2010).

We note that the functions $y_1, \ldots, y_N$ may be projected onto the basis $\theta_1, \ldots, \theta_K$ in order to make the integral in Eq. (6) computable. Alternatively, one may fit model (4) using the raw values of the functions. See Ramsay and Silverman (2005) and Reiss et al. (2010) for further details.

Hypothesis testing

It may often be of interest to assess formally the effect of predictors such as age or diagnostic group on the connectivity–distance profiles. This can be formulated as testing a null model $E[y(s)] = x_0^T \beta_0(s)$ against the alternative model $E[y(s)] = x_1^T \beta_1(s)$ where x_0 is a predictor vector of length m_0 contained within the vector x_1 of length $m_1 > m_0$. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the coefficient function estimates for the two models. Using the notation $\mathrm{SSE}_0(s) = \sum_{i=1}^{N} [y_i(s) - x_{0i}^T \hat{\beta}_0(s)]^2$ and $\mathrm{SSE}_1(s) = \sum_{i=1}^{N} [y_i(s) - x_{1i}^T \hat{\beta}_1(s)]^2$, we can define a pointwise F-statistic at s by

$$F(s) = \frac{[\mathrm{SSE}_0(s) - \mathrm{SSE}_1(s)] / (m_1 - m_0)}{\mathrm{SSE}_1(s) / (N - m_1)} \tag{8}$$

(Ramsay and Silverman, 2005). We can test $F(s)$ at all distances simultaneously, i.e., in a manner that takes into account the multiple s values being tested, by a permutation testing approach (Nichols and Holmes, 2001). By randomly generating many data sets in which the predictor values are permuted, we simulate the null distribution of $\sup_{s \in S} F(s)$. The null model is then rejected at the $100\alpha\%$ level if, for some s, $F(s)$ exceeds the $100(1-\alpha)$ percentile of the permuted-data values of $\sup_{s \in S} F(s)$. Note that when this procedure rejects the null hypothesis, it also identifies the distances s at which the evidence against the null is strong.

Model selection

In the hypothesis testing framework, strong evidence against a simpler (null) model is required in order to reject it in favor of a more complex model. However, hypothesis testing may not always be the most appropriate inferential paradigm. For example, the above permutation test procedure may find each of several scalar predictors to be significant, but we may be interested in knowing which of these is the most predictive of the connectivity–distance profile. More generally, whenever the candidate models are not all nested in each other, we may wish to select the model that optimizes a goodness-of-fit criterion. A natural criterion to use is the cross-validated integrated square error [Eq. (7)].

The relative usefulness of different candidate models may vary as a function of the distance s from the seed. Thus, as an alternative to choosing the model minimizing Eq. (7) as the globally "best" model, Reiss et al. (2010) propose to perform *pointwise model selection* at each s, by comparing the candidate models in terms of the pointwise SSE

$$\sum_{i=1}^{N}\left[y_i(s)-\hat{y}_i^{(-i)}(s)\right]^2, \qquad (9)$$

i.e., the integrand in Eq. (7).

Data acquisition and processing

Image acquisition

The functional imaging data were collected at four sites contributing to the 1000 Functional Connectomes Project (http://www.nitrc.org/projects/fcon_1000). Data sets and associated scan parameters are available at the above URL, where the sites are referred to as Bangor (20 participants), Orangeburg (20), NewYork_a (59), and Newark (19); we refer to them below as sites A–D, respectively.

Image preprocessing

All resting state scans were preprocessed using both AFNI (Cox, 1996) and FSL (www.fmrib.ox.ac.uk). After discarding the first five time points of every scan, the data were corrected for motion by aligning each volume to the mean image using Fourier interpolation in AFNI. Next, the data were spatially smoothed using a 6 mm FWHM Gaussian kernel, and mean-based intensity normalization was performed by scaling all volumes by the same factor. The data were then temporally filtered using both a high-pass filter (Gaussian-weighted least-squares straight line fitting, with $\sigma = 100.0$ s) and a low-pass Gaussian filter (FWHM 2.8 s), followed by linear detrending to remove any residual drift.

Registration

Each individual's high-resolution anatomical image was registered to a common stereotactic space (the Montreal Neurological Institute 152-brain template [MNI152]; 3 mm^3 resolution) using a 12-degrees-of-freedom linear affine transformation implemented in FMRIB's Linear Image Registration Tool (FLIRT), an FSL tool (Jenkinson and Smith, 2001; Jenkinson et al., 2002). After aligning each individual's fMRI data to the corresponding high-resolution anatomical image, the obtained affine transformation was applied to register the fMRI data to the common stereotactic space.

Functional connectivity: seed-based correlation analysis

Consistent with common practice in the resting state functional imaging literature (e.g., Biswal et al., 2010), nuisance signals were regressed out from the data prior to RSFC analyses, to control for the effects of physiological processes such as motion, cardiac and respiratory cycles. Specifically, we regressed each individual's 4D data on the following nine predictors: white matter (WM), cerebrospinal fluid (CSF), the global signal (average time series across all brain voxels), and six motion parameters (movement in the x, y and z directions and rotational movement around 3 axes). To generate the WM and CSF covariates, we segmented each individual's high-resolution structural image using FMRIB's Automated Segmentation Tool (FAST) in FSL, and thresholded the resulting segmented WM and CSF images to obtain masks of ≥80% tissue type probability. Each individual's WM and CSF time series were then calculated by averaging across voxels in the respective masks.

We created six 7.5 mm-radius (33-voxel) seed ROIs centered on the coordinates previously employed by Fox et al. (2005) to examine RSFC for each of six regions, three within the "task-positive" network and three within the "task-negative" network. (These terms refer to brain regions that are more active or less active, respectively, during goal-directed tasks than during rest.) The ROIs within the task-positive network were located in the intraparietal sulcus (IPS; -25, -57, 46), the right frontal eye field (FEF) region of the precentral sulcus (25, -13, 50), and the middle temporal region (MT$^+$; -45, -69, -2). The task-negative network seed ROIs were located in left lateral parietal cortex (LP; -45, -67, 36), medial prefrontal cortex (MPF; -1, 47, -4), and posterior cingulate/precuneus (PCC; -5, -49, 40). Results for the LP seed are given below; plots corresponding to Figs. 3 and 4, for the other five seeds, are provided as Supplementary material.

We extracted the seed time series by averaging across all voxels in each seed ROI from each individual's residual 4D data in MNI152 standard space. For each individual, we determined the correlation between the seed region's time series and that of each voxel in the brain (67,748 voxels in all, excluding the center of the seed), using the AFNI program 3dFIM+in native functional space. Finally, the resulting individual-level correlation maps were converted to z-value maps using Fisher's z transformation.

Application of the proposed methods to the four sites' data

Given the differences among the four sites with respect to data acquisition protocol, scanner, and participant population, we assessed (a) whether the sites differ with respect to LP connectivity as a function of distance, and (b) whether any such differences might be accounted for by sample characteristics, in particular age. After calculating 95th- and 5th-percentile connectivity–distance profiles for each participant, we compared five function-on-scalar regression models: (i) the null model (i.e., estimating the mean connectivity profile only), and regressions on (ii) site, (iii) age, (iv) site and age, and (v) site, age, and site-by-age interaction. The functional data analysis methods described above were implemented in R version 2.9.0 (R Development Core Team, 2010). Nonparametric quantile regression was performed with the R package `quantreg` (Koenker, 2009), and the resulting individual-specific profiles were converted into functional data objects using the R package `fda` (Ramsay et al., 2009). More specifically, these profiles were projected onto a cubic B-spline basis with 25 equally spaced knots spanning the range of distances from the seed, yielding basis dimension $K = 27$ (see de Boor, 2001; Ramsay and Silverman, 2005, for details of B-spline basis construction). Functional linear modeling was implemented with the R package `refund` (Reiss and Huang, 2010), available at http://cran.r-project.org/web/packages/refund or from the authors.

Results

Fig. 2 shows the connectivity–distance profiles for both high and low connectivity at the four sites, with respect to the LP seed. Overall, site A's profiles are the most extreme (highest for high connectivity, and lowest for low connectivity), whereas site B's profiles are the least extreme. However, the boxplots indicate that site A had the youngest participant pool, whereas site B had the oldest. This raises the question as to whether the apparent between-site differences reflect imaging-related factors (e.g., different scan parameters or magnet

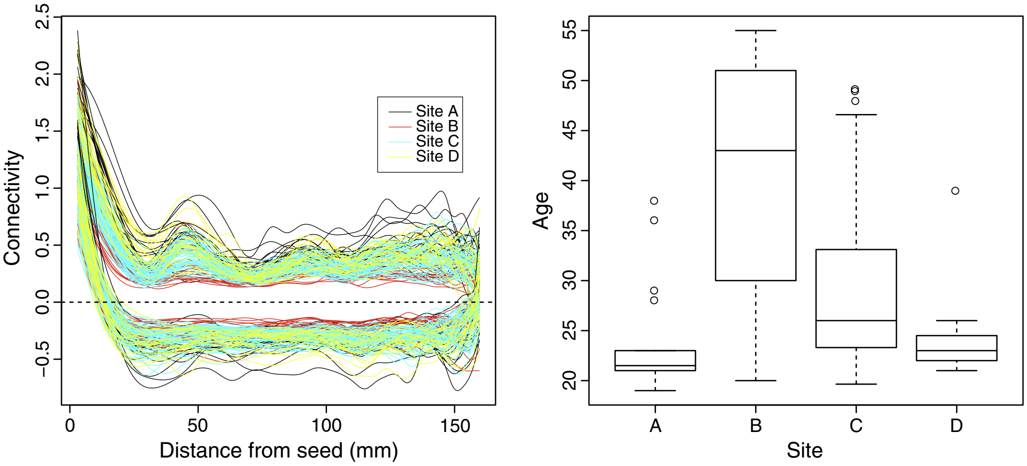

Fig. 2. Left: Each of the upper set of curves represents the high-connectivity (95th percentile) profile for an individual participant with respect to the LP seed; the lower curves are the individual-specific low-connectivity (5th percentile) profiles. The color coding for the four sites suggests systematic differences among them. Right: site-specific age distributions.

strength), or are primarily attributable to age disparities among the four sites.

To help answer this question, we applied our model selection approach to models (i)–(v) referred to above. For each of the four non-null models, Fig. 3 displays the ratio of the pointwise cross-validation score [Eq. (9)] to the score for the null model. All curves lie mostly below 1, indicating that all four models have some predictive value at most distances from the seed. Overall, the site-and-age main effects model appears best, with the site-only model a close second. This suggests systematic differences among the four sites, as well as possible age effects at some distances from the seed. The relatively poor performance of the interaction model suggests that age effects are quite consistent across sites.

Further insight can be gained from Fig. 4, which presents estimates from the regression on site and age [model (iv)]. The most striking between-site differences – in particular, the mean for site A lies above the overall average, whereas that for site B lies below – occur for the high-connectivity profiles at short distances. This is consistent with the observation that, in the left panel of Fig. 3, the lowest portions of the curves for all four non-null models occur at the shortest distances from the LP seed ROI. Note, however, that the confidence intervals are pointwise, i.e., they do not correct for the multiple (in theory, infinitely many) points along the functions. The age effect estimates suggest that high-connectivity profiles tend to decrease with age in the regions below 20 mm and around 50 mm.

We next consider simultaneous testing, to deal with the multiple testing issues not taken into account in Fig. 4. Results of permutation tests for age and site effects on the high-connectivity profiles are shown in Fig. 5. As the two upper plots show, when tested against the null model, both site and age are found to be significant at almost all distances from the seed. These plots provide little guidance as to whether the between-site RSFC differences are in fact due to age disparities. However, when testing age controlling for site and vice versa – in other words, testing model (iv) against model (ii) or model (iii) – the site effect remains intact, whereas the age effect appears significant only in two distance ranges, for short distances and (marginally) around 50 mm from the seed, corresponding to the two main troughs in the lower left plot of Fig. 4.

These results suggest that while the shape of the connectivity–distance profiles tends to be similar across sites, there are important between-site differences. These differences may be attributable in

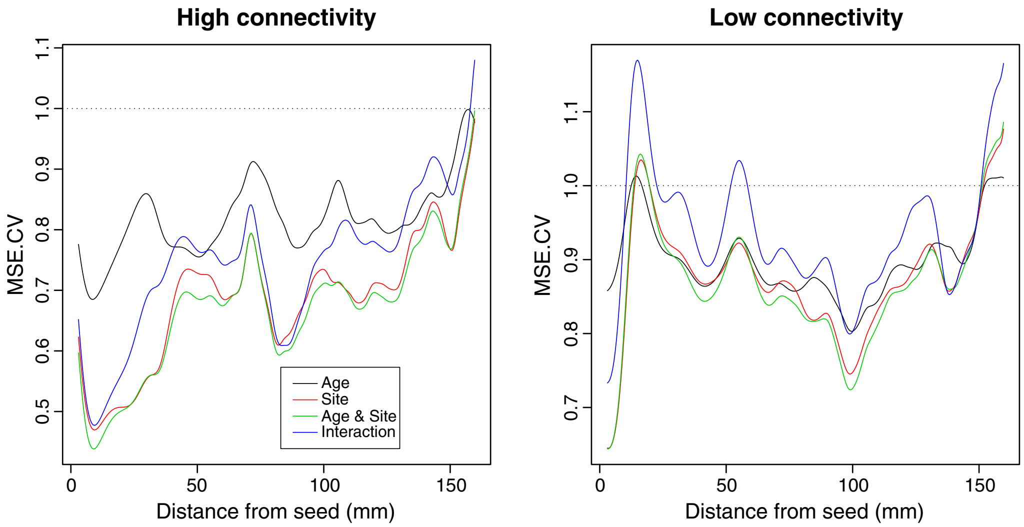

Fig. 3. Pointwise cross-validation scores for four LP seed models (high-connectivity profiles at left, low-connectivity at right), expressed as a ratio with respect to score for the null model (i.e., values below 1 indicate a better fit than the null model).

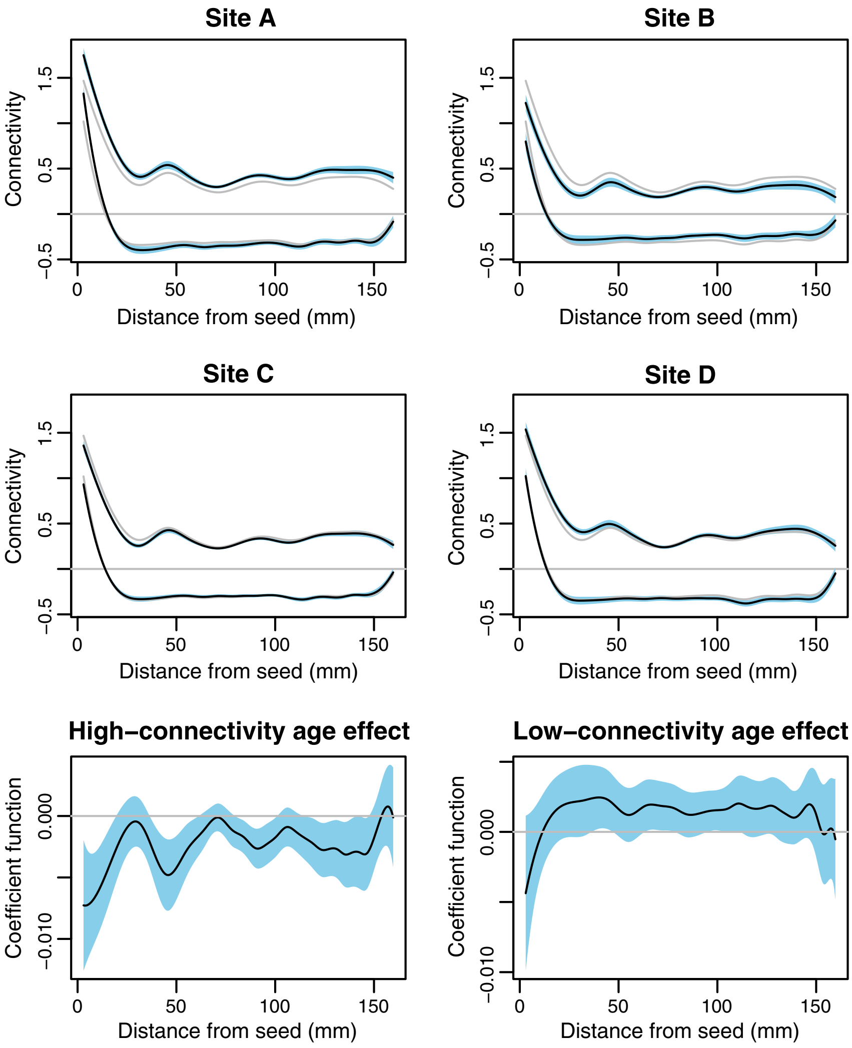

Fig. 4. Estimates obtained by regressing the high- and low-connectivity profiles (for the LP seed) on site and age. First four plots: mean functions (overall mean plus site effect) for the four sites (black curves, with confidence intervals shown in blue). The grey curves represent overall mean functions for the entire sample. Last two plots: age effect estimates and confidence intervals.

part to the different magnets. For instance, site B's 1.5-T magnet has inherently lower signal-to-noise ratio than those at the other sites, which likely resulted in attenuated correlations with the seed. This could help to explain the relatively low magnitude of site B's profiles.

Fig. 6 provides an alternative view of the age effects on both high- and low-connectivity profiles, controlling for site. The green and blue lines are expected profiles (for an "average" site) at the first and third quartiles of the pooled age distribution (ages 23 and 35, respectively). Yellow shading indicates distances at which the permutation tests for high connectivity (see the lower left plot in Fig. 5) or low connectivity (not shown) found age effects significant at the 5% level. Excluding very short distances, the clearest discrepancy between the blue and green curves (i.e., age effect) is around 50 mm from the LP seed (the corresponding peak in the lower left plot in Fig. 5 is at the threshold of permutation test significance). This marked discrepancy coincides with a major peak in both expected high-connectivity profiles. More generally, these data suggest that with increasing age, the expected profiles for both high and low RSFC tend to be attenuated across the range of distances, in particular at peaks. This age effect (cf. the age effects in Fig. 4, and parallel results in the Supplementary material, especially for the PCC seed) may suggest a continuation, from early to mid-adulthood, of the shift toward a more distributed RSFC architecture that occurs from childhood to adulthood (Fair et al., 2007, 2009; Kelly et al., 2009).

It is interesting to contrast Fig. 6 with Fig. 7, which displays thresholded z-statistic maps produced by using the general linear model (GLM) implemented in FSL to perform a voxelwise group analysis of RSFC. The red and blue regions are those in which connectivity with the seed is significantly positive or negative, respectively ($p<.05$, corrected by Gaussian random field theory). Yellow and green denote positive and negative age effects. The negative age effect cluster lies in a region of positive connectivity immediately around the seed, consistent with the negative age effect at short distances in Fig. 6. Although our method provides information about the tails of the RSFC distribution within a sphere of any given radius centered at the seed, it cannot pinpoint the location of clusters of significant voxels. Moreover, our method does not find a significant positive age effect at long distances corresponding to the clusters shown in yellow in Fig. 7. On the other hand, Fig. 6 reveals at a glance the overall pattern that gives rise to these clusters, namely, an attenuation of negative connectivities at all distances from the seed. This global pattern cannot be discerned from Fig. 7, but for some purposes may be more important than particular suprathreshold clusters. For instance, overall patterns of this kind might be helpful for succinctly comparing the effect of a covariate on different seeds' connectivity maps.

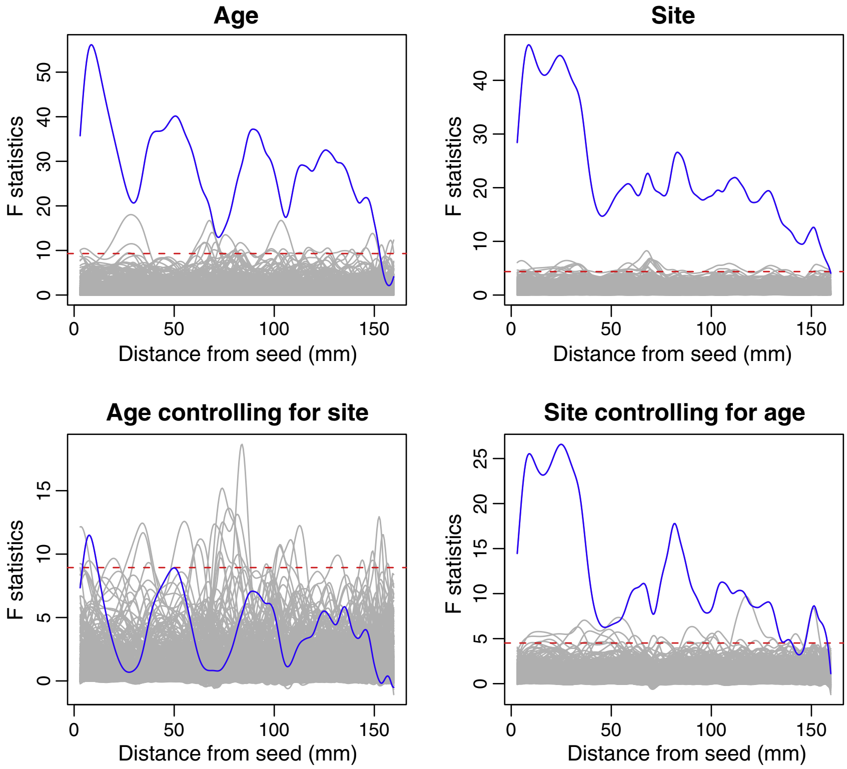

Fig. 5. Permutation tests for effects of age and site on the high-connectivity profiles. The two upper plots refer to models with age alone or site alone. The blue curves give *F*-statistics [Eq. (8)] at each distance from the LP seed; the grey curves do the same for each of 499 permuted data sets; and the red dashed line indicates the threshold for significance at the 5% level. The two lower plots are similar, but refer to testing each predictor with the other predictor in the model.

Discussion

We have introduced a straightforward methodology that uses nonparametric quantile regression to summarize whole-brain RSFC with a seed as a function of distance, and regresses functions of this kind on subject-specific predictors. Our example application, using a pool of participants scanned at four different sites, demonstrates the utility of this approach for comparing groups and studying covariate effects. In particular, our analysis (a) confirmed the existence of site effects over and above age differences among the four samples (Fig. 5), and (b) revealed a general tendency toward less extreme connectivities with increasing age (Fig. 6).

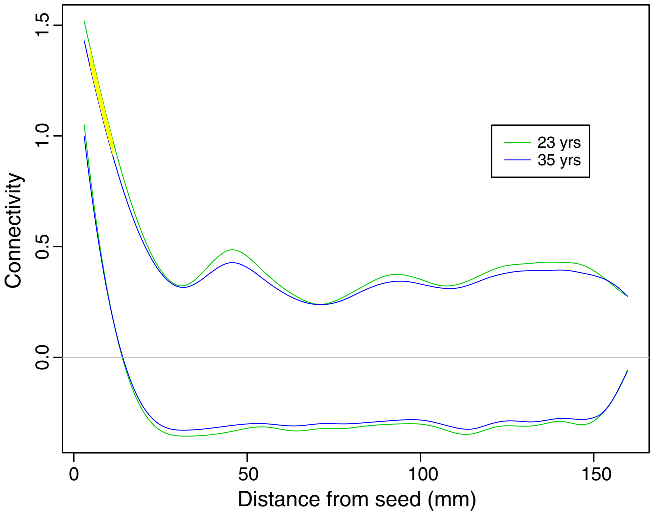

Fig. 6. Expected high- and low-connectivity profiles at the first and third quartiles of the age distribution for the four sites. Yellow shading indicates distances from the LP seed at which the permutation tests (displayed at lower left in Fig. 5 for high connectivity) found significant age effects.

The connectivity profiles we have proposed are reminiscent of Figs. 2 and 6 of Andrews-Hanna et al. (2007), which present average *z*-transformed correlations for two age groups along a line from a seed region to a prespecified "target region." Considering only voxels along a particular line segment obviates the need for the quantile regression procedures that we employ. However, our approach does not require specifying a target region, and thus allows for exploration of whole-brain RSFC with the given seed.

In addition to their role in understanding how different sources of variation affect seed-based RSFC, connectivity–distance profiles can prove very useful for quality assurance. For example, while preparing the analyses presented above, we initially found one participant's high-connectivity profile to differ markedly from the rest of the sample. Further investigation traced this anomaly to a preprocessing error that otherwise might have gone undetected.

Our statistical methodology may be extended in several ways. Functional mixed effects models (e.g., Guo, 2002) can handle multiple sources of random variation; in our case, each individual's measured profile would be modeled as a sum of appropriate fixed effect curves plus a random effect curve plus measurement error. Such models are more computationally intensive than the fixed effects model pursued here, especially for large numbers of profiles (but see Greven et al., 2010, for recent advances). Another possible avenue for further research would be to combine the nonparametric quantile regression and the functional linear model into a single step. This is the approach of varying-coefficient quantile regression models (e.g., Cai and Xu, 2008), and it would be interesting to see if such models are computationally feasible with tens of thousands of data points per individual.

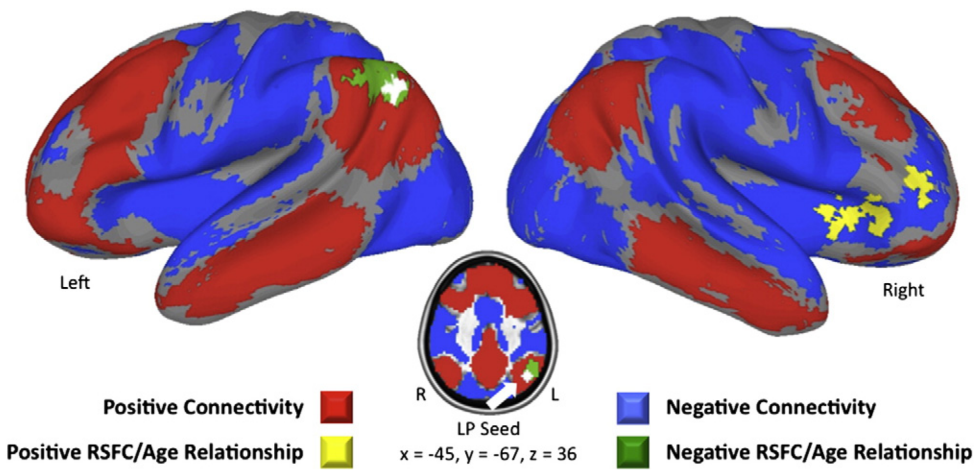

Fig. 7. Regions of significant positive (red) and negative (blue) connectivity with the LP seed (white), and regions exhibiting positive (yellow) or negative (green) effects of age on RSFC with the seed, based on a GLM implemented through FSL FEAT (ordinary least squares; $p<.05$, corrected by Gaussian random field theory). The model included site and age, as well as gender, as covariates.

One limitation of our method is that if several regions of roughly equal distance are highly correlated with the seed, but the connectivity of only one of these regions is related to a predictor such as age, our model might fail to detect this effect. If an *a priori* hypothesis, or the investigators' interest, focuses on voxels within a geometrically or anatomically defined portion of the brain, one can make the analysis more sensitive by applying a mask to retain only those voxels.

Several choices were made in defining the seed-based RSFC maps on which the connectivity–distance profiles were based. In accordance with common practice, RSFC was defined as simple correlation, and the seed time series was computed by averaging over all voxels within the seed region. It should be noted that other measures of time series similarity are available (Liao, 2005), and that some authors use the first principal component to derive the seed time series (Boly et al., 2008). We regressed out the whole-brain or global signal prior to computing the resting state correlations, yet there has been much debate about the meaning of the negative correlation estimates that result from this step (e.g., Murphy et al., 2009; Chang and Glover, 2009; Van Dijk et al., 2010). However, none of these choices is fundamental to the proposed functional regression methodology; as such, evaluation of alternatives lies beyond the scope of this paper.

The seed-based RSFC paradigm has been criticized for its reliance on an *a priori* selected region of interest and for its inability to provide information about causal relationships (Boly et al., 2008). The former limitation actually points to a virtue of our approach; applying it to multiple seeds would be a fast and straightforward way to identify seeds of interest, which could subsequently be examined by standard voxelwise group analyses. At any rate, we wish to emphasize that our method has numerous potential applications beyond seed-based connectivity. The horizontal and vertical axes of the functional data can refer to any two voxelwise quantities whose functional relationship is of interest. For example, one might wish to study profiles representing amplitude of low-frequency fluctuations (ALFF; Zuo et al., 2010) as a function of cortical thickness (Hutton et al., 2008), regional homogeneity (Zang et al., 2004), or voxel-based morphometry measures (Ashburner and Friston, 2000).

More generally, as noted above, our model can be thought of as an extension of moderation analysis to a class of more complex questions arising in neuroimaging—questions of the form: Does variable x affect the tendency for voxels with certain values of s to have unusually high values of y? In this paper, s was distance from a seed, y referred to RSFC with that seed, and x was age or site. But any of s, y or x, or all three, might be different in other applications. With the proliferation of ever-richer data sets combining multiple imaging modalities with large collections of clinical and demographic variables, the need to uncover such complex relationships is growing rapidly. We believe, therefore, that the approach presented here could be adapted to many other settings, and could prove to be a powerful tool for enriching our understanding of the brain.

Acknowledgments

The authors wish to express their gratitude to the referees, whose incisive comments led to a much improved paper, and to Clare Kelly, for very helpful discussions. This research was partially supported by grants from the National Institute of Mental Health (R01MH083246 and K23MH087770), Autism Speaks, the Stavros Niarchos Foundation, and the Leon Levy Foundation, and gifts from Joseph P. Healy, Linda and Richard Schaps, Jill and Bob Smith, and the endowment provided by Phyllis Green and Randolph Cōwen. Reiss's research was supported in part by National Science Foundation grant DMS-0907017 and National Institutes of Health (NIH) grant R01 EB009744-01A. Hoptman's research was supported in part by NIH grants R21 MH084031 and R01 MH064783.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.neuroimage.2011.01.071.

References

Andrews-Hanna, J.R., Snyder, A.Z., Vincent, J.L., Lustig, C., Head, D., Raichle, M.E., Buckner, R.L., 2007. Disruption of large-scale brain systems in advanced aging. Neuron 56, 924–935.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. NeuroImage 11, 805–821.

Biswal, B., Yetkin, F.Z., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn. Reson. Med. 34, 537–541.

Biswal, B.B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S.M., Beckmann, C.F., Adelstein, J.S., Buckner, R.L., Colcombe, S., Dogonowski, A.M., Ernst, M., Fair, D., Hampson, M., Hoptman, M.J., Hyde, J.S., Kiviniemi, V.J., Kötter, R., Li, S.J., Lin, C.P., Lowe, M.J., Mackay, C., Madden, D.J., Madsen, K.H., Margulies, D.S., Mayberg, H.S., McMahon, K., Monk, C.S., Mostofsky, S.H., Nagel, B.J., Pekar, J.J., Peltier, S.J., Petersen, S.E., Riedl, V., Rombouts, S.A., Rypma, B., Schlaggar, B.L., Schmidt, S., Seidler, R.D., Siegle, G.J., Sorg, C., Teng, G.J., Veijola, J., Villringer, A., Walter, M., Wang, L., Weng, X.C., Whitfield-Gabrieli, S., Williamson, P., Windischberger, C., Zang, Y.F., Zhang, H.Y., Castellanos, F.X., Milham, M.P., 2010. Toward discovery science of human brain function. Proc. Natl Acad. Sci. USA 107, 4734–4739.

Boly, M., Phillips, C., Tshibanda, L., Vanhaudenhuyse, A., Schabus, M., Dang-Vu, T.T., Moonen, G., Hustinx, R., Maquet, P., Laureys, S., 2008. Intrinsic brain activity in altered states of consciousness: how conscious is the default mode of brain function? Ann. NY Acad. Sci. 1129, 119–129.

Cai, Z., Xu, X., 2008. Nonparametric quantile estimations for dynamic smooth coefficient models. J. Am. Stat. Assoc. 103, 1595–1608.

Chang, C., Glover, G.H., 2009. Effects of model-based physiological noise correction on default mode network anti-correlations and correlations. Neuroimage 47, 1448–1459.

Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29 (3), 162–173.

Damoiseaux, J.S., Rombouts, S.A.R.B., Barkhof, F., Scheltens, P., Stam, C.J., Smith, S.M., Beckmann, C.F., 2006. Consistent resting-state networks across healthy subjects. Proc. Natl Acad. Sci. USA 103, 13848–13853.

de Boor, C., 2001. A Practical Guide to Splines, rev. ed. Springer-Verlag, New York.

Fair, D.A., Dosenbach, N.U.F., Church, J.A., Cohen, A.L., Brahmbhatt, S., Miezin, F.M., Barch, D.M., Raichle, M.E., Petersen, S.E., Schlaggar, B.L., 2007. Development of distinct control networks through segregation and integration. Proc. Natl Acad. Sci. USA 104, 13507–13512.

Fair, D.A., Cohen, A.L., Power, J.D., Dosenbach, N.U.F., Church, J.A., Miezin, F.M., Schlaggar, B.L., Petersen, S.E., 2009. Functional brain networks develop from a "local to distributed" organization. PLoS Comput. Biol. 5 (5), e1000381.

Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E., 2005. The human brain is intrinsically organized into dynamic anticorrelated functional networks. Proc. Natl Acad. Sci. USA 102, 9673–9678.

Greven, S., Crainiceanu, C., Caffo, B., Reich, D., 2010. Longitudinal functional principal component analysis. Electron. J. Stat. 4, 1022–1054.

Guo, W., 2002. Functional mixed effects models. Biometrics 58, 121–128.

Hutton, C., De Vita, E., Ashburner, J., Deichmann, R., Turner, R., 2008. Voxel-based cortical thickness measurements in MRI. Neuroimage 40, 1701–1710.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17 (2), 825–841.

Kelly, A.M.C., Di Martino, A., Uddin, L.Q., Shehzad, Z., Gee, D.G., Reiss, P.T., Margulies, D.S., Castellanos, F.X., Milham, M.P., 2009. Development of anterior cingulate functional connectivity from late childhood to early adulthood. Cereb. Cortex 19, 640–657.

Koenker, R., 2005. Quantile Regression. Cambridge University Press, New York.

Koenker, R., 2009. quantreg: quantile regression. R package version 4.27. http://CRAN.R-project.org/package=quantreg.

Koenker, R., Ng, P., Portnoy, S., 1994. Quantile smoothing splines. Biometrika 81, 673–680.

Liao, T.W., 2005. Clustering of time series data—a survey. Pattern Recognit. 38, 1857–1874.

Margulies, D.S., Kelly, A.M.C., Uddin, L.Q., Biswal, B.B., Castellanos, F.X., Milham, M.P., 2007. Mapping the functional connectivity of anterior cingulate cortex. Neuroimage 37, 579–588.

Murphy, K., Birn, R.M., Handwerker, D.A., Jones, T.B., Bandettini, P.A., 2009. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? Neuroimage 44, 893–905.

Nichols, T.E., Holmes, A.P., 2001. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis, 2nd ed. Springer, New York.

Ramsay, J.O., Hooker, G., Graves, S., 2009. Functional Data Analysis with R and MATLAB. Springer, New York.

Reiss, P.T., Huang, L., 2010. refund: regression with functional data. R package version 0.1-0. URL http://cran.r-project.org/web/packages/refund.

Reiss, P.T., Huang, L., Mennes, M., 2010. Fast function-on-scalar regression with penalized basis expansions. Int. J. Biostat. 6 (1) article 28.

Rice, J.A., Silverman, B.W., 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. J. R. Stat. Soc. B 53, 233–243.

Roy, A.K., Shehzad, Z., Margulies, D.S., Kelly, A.M.C., Uddin, L.Q., Gotimer, K., Biswal, B.B., Castellanos, F.X., Milham, M.P., 2009. Functional connectivity of the human amygdala using resting state fMRI. Neuroimage 45, 614–626.

Ruppert, D., 2002. Selecting the number of knots for penalized splines. J. Comput. Graph. Stat. 11, 735–757.

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, New York.

Van Dijk, K.R.A., Hedden, T., Venkataraman, A., Evans, K.C., Lazar, S.W., Buckner, R.L., 2010. Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. J. Neurophysiol. 103, 297–321.

Viviani, R., Grön, G., Spitzer, M., 2005. Functional principal component analysis of fMRI data. Hum. Brain Mapp. 24, 109–129.

Wood, S.N., 2006. Generalized Additive Models: An Introduction with R. Chapman & Hall, Boca Raton, FL.

Zang, Y., Jiang, T., Lu, Y., He, Y., Tian, L., 2004. Regional homogeneity approach to fMRI data analysis. Neuroimage 22, 394–400.

Zhou, D., Thompson, W.K., Siegle, G., 2009. MATLAB toolbox for functional connectivity. Neuroimage 47, 1590–1607.

Zhu, H., Styner, M., Tang, N., Liu, Z., Lin, W., Gilmore, J.H., 2010. FRATS: functional regression analysis of DTI tract statistics. IEEE Trans. Med. Imaging 29, 1039–1049.

Zuo, X.-N., Di Martino, A., Kelly, C., Shehzad, Z.E., Gee, D.G., Klein, D.F., Castellanos, F.X., Biswal, B.B., Milham, M.P., 2010. The oscillating brain: complex and reliable. Neuroimage 49, 1432–1445.